Predictive Modeling in Marketing

To develop scientific method of communication for the next planned campaigns

May 2018

Background

Client: A leading FMCG company of India

Objective:

1. Is to learn from campaign responses and develop a statistical model to identify which customers should be targeted for the next campaign. The model can be used to select targeted base for the next campaign
2. Identify most effective communication channel

Data available for the analysis :

1. Transactions of the customers two years prior to the campaign
2. Communication channels of the previous campaign
3. Buying behaviour ,level of engagement and response of the customers to the previous campaign
4. Master file for Regions

Key Steps In Model Building

Data Management

Understanding and preprocessing

Exploratory Data Analysis

Understand some interesting patterns in the data through visualization

Develop The Statistical Model

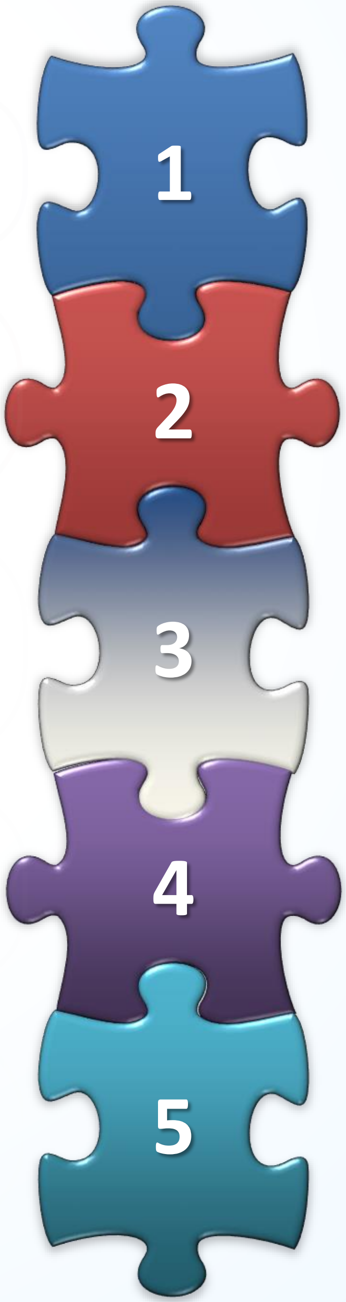
Develop a model using appropriate Dependent and Independent variables

Model Validation

Validate the model using Hold Out and K-fold Cross validation methods. Derive Area under the ROC curve for each model

Model Implementation

Implement the model using significant predictors and calculate predicted probabilities



Data Highlights

- Total number of customers for which region is recorded: 90,000
- Total transactions recorded: 5,00,000
- Number of brands for which transactions were recorded: 7

The campaign run in January 2015 for SKU in Brand 1

- Sample size with unique customer ID’s : 1,228
- Response to the campaign run in January 2015 was recorded as a binary variable

1 = Responded   0 = Did not respond

- Response rate ≈ 40%

Data Files Snapshots

1.Transaction Details

Customer	Date	Month	Year	Brand	Sales
10000	5/20/2014	5	2014	B4	21793
10000	10/24/2014	10	2014	B5	7155
10000	08-01-2014	8	2014	B1	29630
10000	10/20/2014	10	2014	B3	1530
10000	01-11-2013	1	2013	B2	3965
10000	4/19/2013	4	2013	B2	34608
10001	3/15/2014	3	2014	B2	39256
10001	10/29/2013	10	2013	B5	14612
10001	12/16/2014	12	2014	B2	2902
10001	07-05-2014	7	2014	B1	6122
10001	6/14/2014	6	2014	B1	20355
10002	12/19/2013	12	2013	B4	6468
10002	10-05-2013	10	2013	B5	36800
10002	05-07-2013	5	2013	B1	6649
10003	07-09-2013	7	2013	B4	21076
10003	03-06-2013	3	2013	B5	6768
10004	12-08-2013	12	2013	B4	32573
10004	8/30/2014	8	2014	B5	34218
10004	11-11-2014	11	2014	B1	6783

2. Campaign Response Details

Customer	response	n_comp	loyalty	portal	rewards	nps	n_yrs
18263	1	2	0	1	0	7	8
50429	0	1	1	1	1	3	3
98593	1	0	1	0	0	9	6
44804	0	4	1	1	1	2	5
81015	0	4	1	1	1	2	2
15273	1	2	1	1	1	5	7
51484	1	1	0	0	0	6	6
87695	0	3	0	1	0	8	3
33906	0	3	1	0	1	2	3
70117	0	5	1	1	1	8	6
73807	0	4	1	1	1	0	8
47262	1	4	1	1	1	4	2
99997	1	5	1	0	1	2	8

3. Communication Channels

Customer	email	sms	call
10048	1	0	0
10073	1	0	1
10258	1	0	0
10416	1	0	1
10444	0	0	1
10454	0	1	0
10512	0	1	0
10618	1	0	1
10653	1	0	1
10819	1	1	1
10831	2	1	3
10836	1	1	1
10869	3	2	1

4. Master file for Regions

Customer	Region
10000	North
10001	South
10002	West
10003	South
10004	East
10005	West
10006	West
10007	South
10008	East
10009	North
10010	South
10011	East
10012	East
10013	South

Data Management

Data Understanding and Pre-processing

- **Basics About the Data**: Understanding data dimensions, variable types, variable relationships

- **Identifying Modeling variables :** Response to the campaign run in January 2015 was considered as the dependent variable. Independent variables were identified based on business understanding.

- **Converting Raw Data to Usable Data :**

1. Checking for and handling :

- ❑ Missing Values
- ❑ Inconsistencies

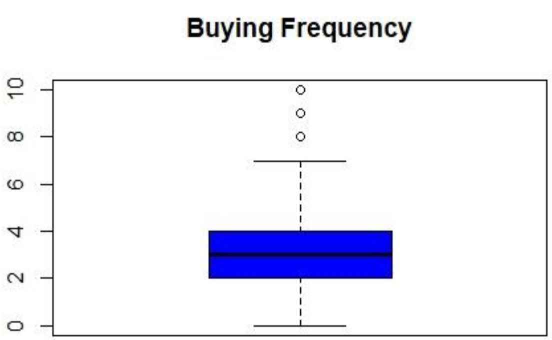
2. Independent Variables were not directly available and were derived from different datasets. All these newly derived variables were compiled in Master file for further analysis.

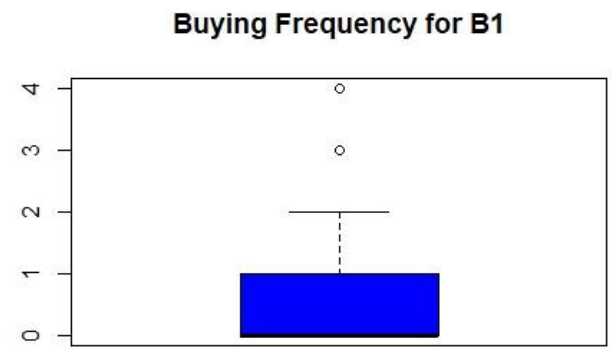
- **Pre-Processing:**

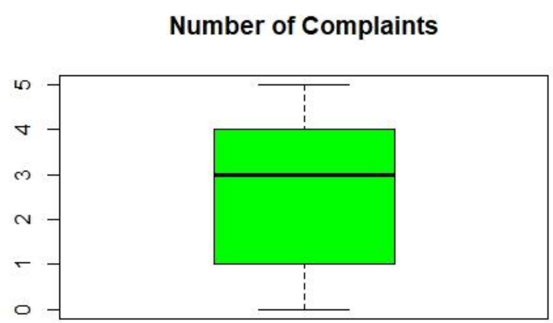
1. Grouping/ Merging of 4 Data files

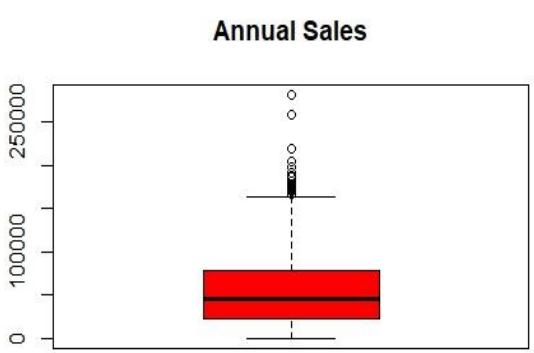
Exploratory Data Analysis

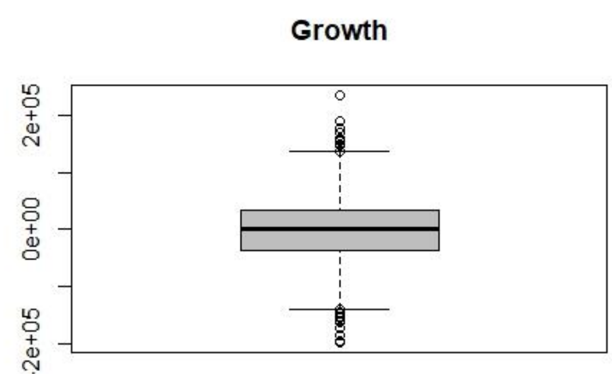
Explore Patterns Using Box-plots



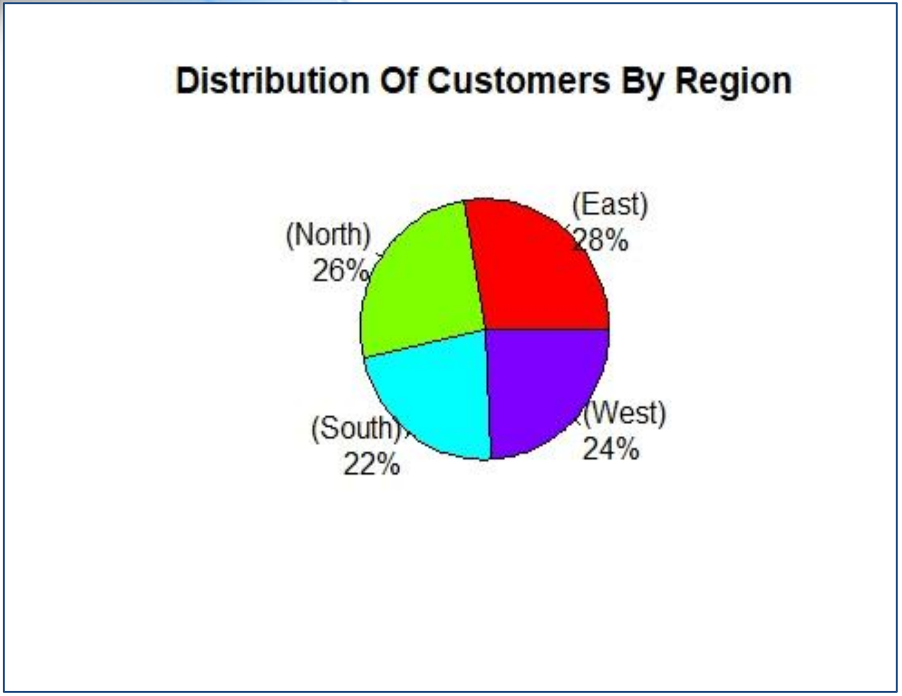


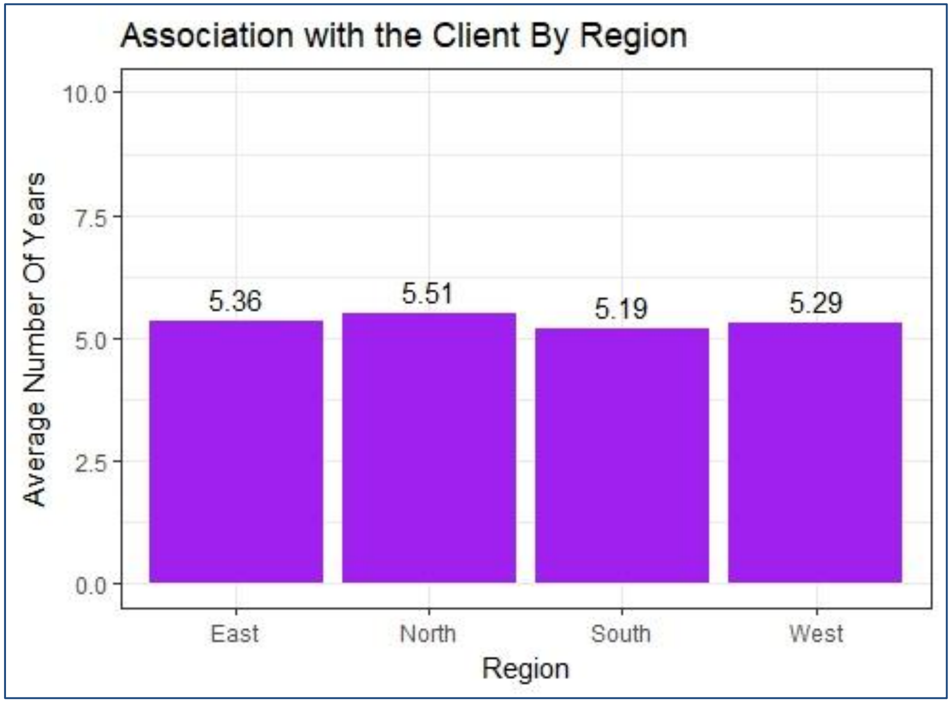






Distribution of Customers by Region



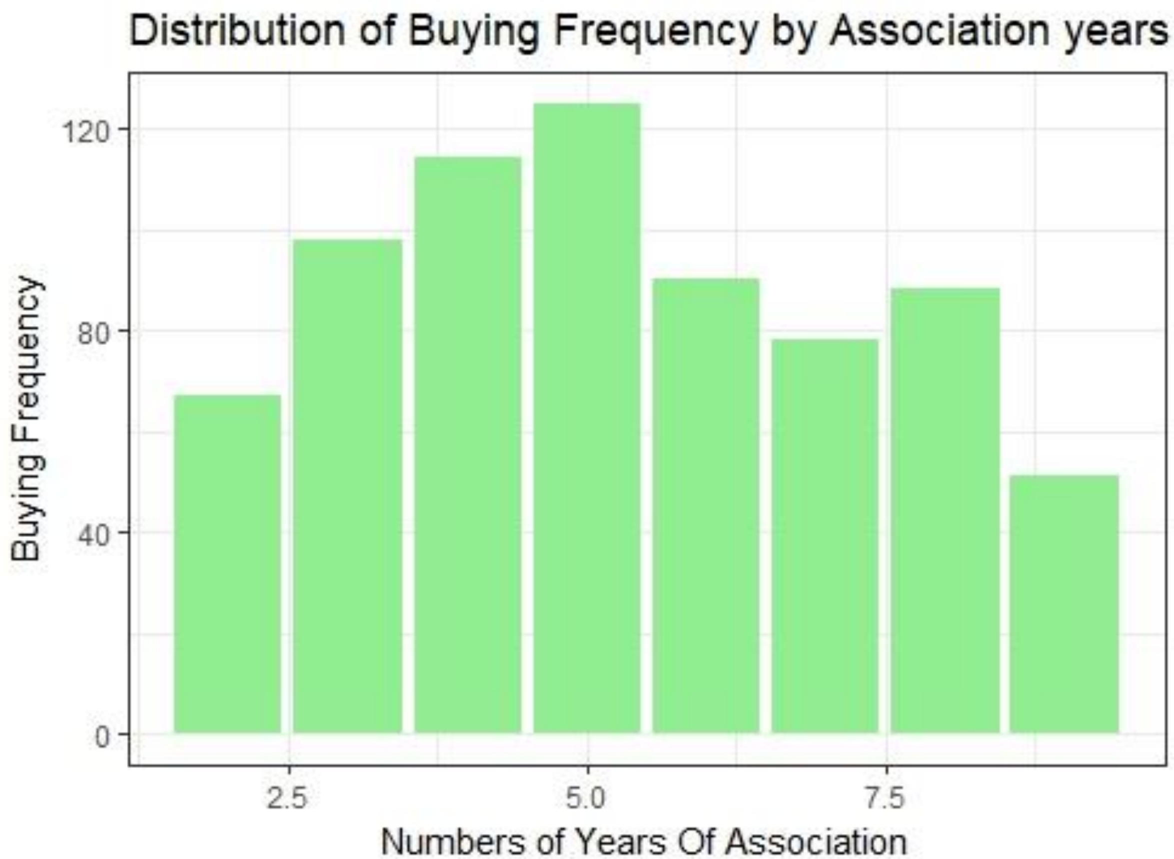


*East Region contributes highest number of customers in the Data followed by North Region.

*On average North Region has highest association years with the client

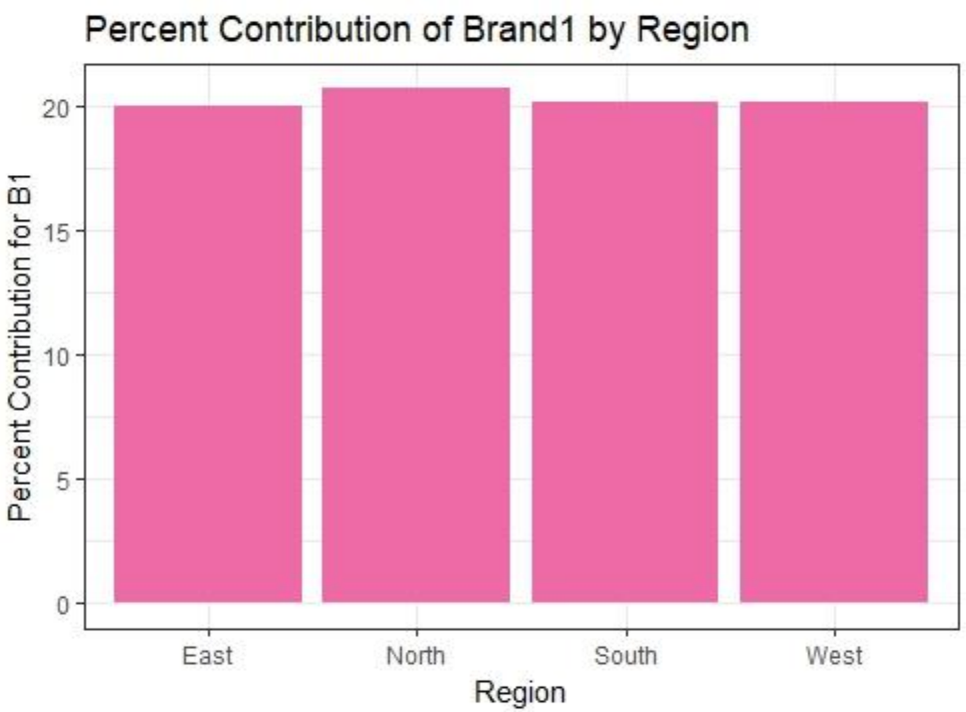
Distribution of Buying Frequency Of Brand1 By Number of Association Years

n_yrs	buyingfreq_B1
2	67
3	98
4	114
5	125
6	90
7	78
8	88
9	51



*Buying Frequency for Brand1 seems to be higher for the customers who have 4 and 5 years of association. On the contrary, it is lower in the initial years of collaboration.

Region wise Contribution for Brand1



*Region wise contribution for Brand1 varies with a slight difference. North region shows a highest contribution of approximately 21% compared to other regions

Distribution of the different Communication Channels by Number of Customers

master$sms	master$response 0	1	Row Total
0	253	170	423
	0.598	0.402	0.344
1	478	282	760
	0.629	0.371	0.619
2	19	20	39
	0.487	0.513	0.032
3	0	3	3
	0.000	1.000	0.002
4	1	2	3
	0.333	0.667	0.002
Column Total	751	477	1228
	0.612	0.388	

master$call	master$response 0	1	Row Total
0	379	158	537
	0.706	0.294	0.437
1	369	271	640
	0.577	0.423	0.521
2	3	36	39
	0.077	0.923	0.032
3	0	10	10
	0.000	1.000	0.008
4	0	1	1
	0.000	1.000	0.001
5	0	1	1
	0.000	1.000	0.001
Column Total	751	477	1228
	0.612	0.388	

master$email	master$response 0	1	Row Total
0	317	107	424
	0.748	0.252	0.345
1	418	328	746
	0.560	0.440	0.607
2	16	40	56
	0.286	0.714	0.046
3	0	2	2
	0.000	1.000	0.002
Column Total	751	477	1228
	0.612	0.388	

The above 3 tables shows the distribution of communication channels and the appropriate response given. The results shows significant rise in the response after 2 follows ups and surprisingly this is consistent for each of the 3 communication channels. So ,we can conclude that more follow up with the customers is the key to gain the positive response in the campaign.

Distribution of Communication Channels

	sms	call	email	response	Total Number Responded to the Campaign	Total Number Targetted	Proportion
6	1	2	0	1	1	1	100%
7	2	2	0	1	1	1	100%
13	2	1	1	1	5	5	100%
14	3	1	1	1	1	1	100%
15	4	1	1	1	1	1	100%
16	0	2	1	1	7	7	100%
17	1	2	1	1	6	6	100%
18	2	2	1	1	1	1	100%
19	0	3	1	1	1	1	100%
20	1	3	1	1	1	1	100%
24	3	0	2	1	1	1	100%
25	4	0	2	1	1	1	100%
27	1	2	2	1	9	9	100%
28	2	2	2	1	1	1	100%
29	3	2	2	1	1	1	100%
30	0	3	2	1	2	2	100%
31	1	3	2	1	6	6	100%
32	1	4	2	1	1	1	100%
33	0	5	2	1	1	1	100%
34	2	0	3	1	1	1	100%
35	2	1	3	1	1	1	100%
10	2	0	1	1	5	6	83%
22	1	0	2	1	8	10	80%
26	0	2	2	1	4	5	80%
9	1	0	1	1	19	26	73%
4	1	1	0	1	5	7	71%
5	0	2	0	1	5	7	71%
23	2	0	2	1	1	2	50%
12	1	1	1	1	137	320	43%
11	0	1	1	1	119	291	41%
8	0	0	1	1	25	80	31%
21	0	0	2	1	4	13	31%
1	1	0	0	1	89	373	24%
2	2	0	0	1	4	21	19%
3	0	1	0	1	2	11	18%

To find number of SMS, calls and email ,the Frequency table was generated by aggregating columns SMS, email , calls & Response by Customer ID to understand the distribution of the communication channels in the data . The cases for which there was no Call, SMS or email sent were further investigated to see what was the response given in the campaign.

Similarly , the cases for which at least any one of communication channel was recorded, the response variable was checked in the data.

Distribution of Communication Channels

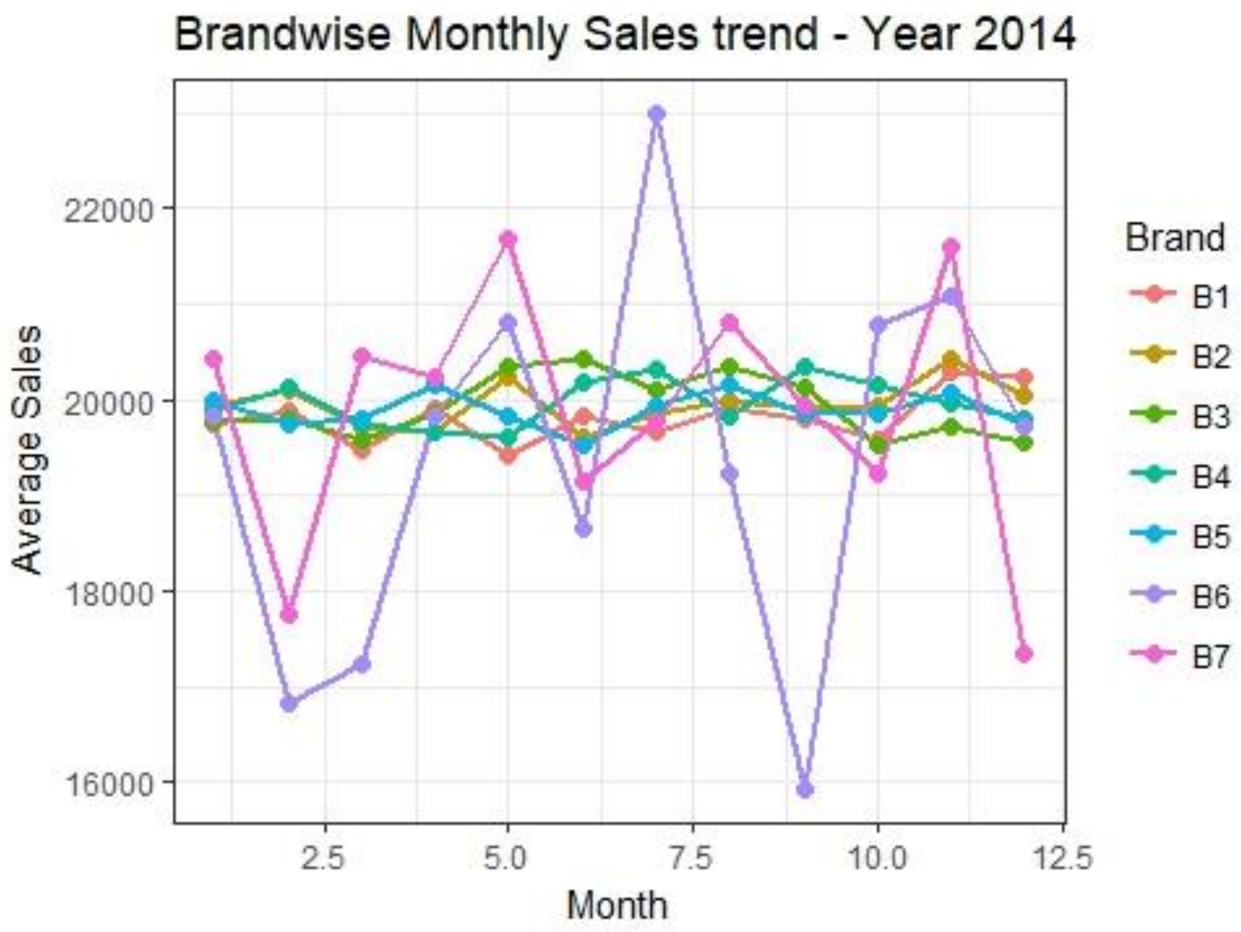
Proportion for the response given by the Customers for at least one of the Communication channel mentioned

Response	Count	Proportion
0	749	61%
1	477	39%

There are 2 Customer ID's ,not invited for the campaign. We can obviously expect no response for them.

Customer ID	Response
27019	0
31257	0

Brand wise Monthly Sales Trend- Year 2014



Approach

Methods Used

- Binary Logistic Regression
- Random Forests
- Naïve Bayes Classifier
- Support Vector Machines

Binary Logistic Regression

Dependent Variable : Response

Independent Variables :

1. Annual Sales
2. Sales in Q4'14
3. Sales in Brand1
4. % contribution in Brand1
5. Association with the Client(number of years)
6. Buying Frequency
7. Buying Frequency of Brand 1
8. Region
9. Net Promotion Score
10. Loyalty
11. Portal Membership
12. Satisfaction Levels(Number of complaints)
13. Communication Channels
14. Brand Engagement
15. Growth

Logistic Regression in R :Model Output

Coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.75E+00	3.43E-01	-8.021	1.05e-15 ***
Annualsales	1.76E-06	3.30E-06	0.533	0.5938
sales_Q4	-4.00E-06	3.80E-06	-1.053	0.29249
Sales_B1	5.51E-07	6.97E-06	0.079	0.93703
Pcontri_B1	-1.55E-03	3.76E-03	-0.412	0.68064
n_yrs	8.71E-02	3.14E-02	2.773	0.00555 **
buyingfreq	1.29E-01	9.67E-02	1.33	0.18344
buyingfreq_B1	9.65E-02	1.66E-01	0.581	0.56105
RegionNorth	-3.32E-02	1.77E-01	-0.188	0.85117
RegionSouth	-2.40E-02	1.83E-01	-0.131	0.89582
RegionWest	9.75E-02	1.77E-01	0.549	0.5827
nps	1.60E-01	2.47E-02	6.494	8.36e-11 ***
loyalty	-4.79E-01	6.63E-01	-0.722	0.47012
portal	-4.43E-02	1.64E-01	-0.271	0.78636
n_comp	-2.92E-02	4.37E-02	-0.669	0.50375
email	9.37E-01	1.70E-01	5.505	3.69e-08 ***
sms	4.33E-01	1.55E-01	2.793	0.00523 **
Call	1.39E+00	2.34E-01	5.943	2.80e-09 ***
rewards	-8.50E-01	6.88E-01	-1.235	0.21679
brandengagement	-1.76E-01	1.13E-01	-1.556	0.11983
growth	-2.71E-06	1.47E-06	-1.846	0.06482 .

n_yrs, nps, email, sms, call are statistically significant

Logistic Regression in R :Model Output

Logistic Model using only significant predictors

```
fmcg_glmmodel <- glm(response ~n_yrs+nps+email+sms+call,data=Master_FMCG,family=binomial)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6123  -0.9514  -0.6457   1.1172   2.1122

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.80301    0.25338 -11.063  < 2e-16 ***
n_yrs        0.09106    0.03064   2.972 0.002958 **
nps          0.15765    0.02390   6.597 4.21e-11 ***
email        0.63788    0.14781   4.315 1.59e-05 ***
sms          0.41289    0.11754   3.513 0.000444 ***
call         0.62035    0.13540   4.582 4.62e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1640.7  on 1227  degrees of freedom
Residual deviance: 1481.0  on 1222  degrees of freedom
AIC: 1493

Number of Fisher Scoring iterations: 4
```

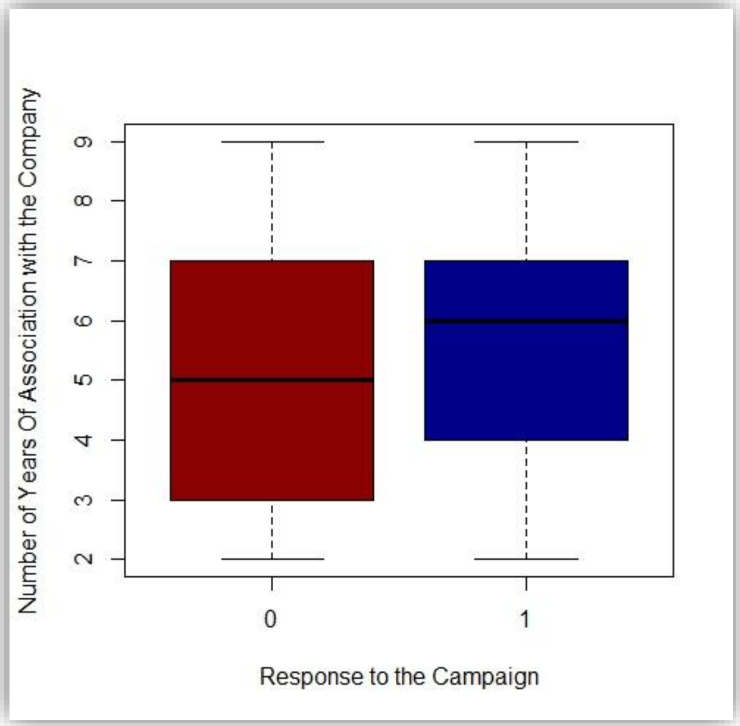
Model Equation:

response= -2.80+0.0910(n_yrs)+0.1577(nps)+ 0.6379(email)+0.4129(sms)+0.6203(call)

Visualizing Distribution Graphically For Significant Predictors

Number of Years Of Association by Response to the Campaign

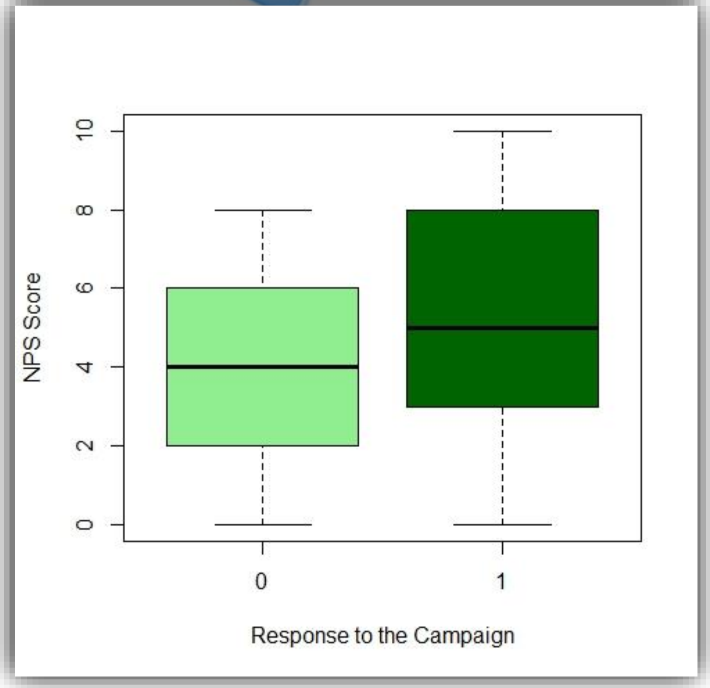
Boxplot for Number of Years of Association By Response to the Campaign



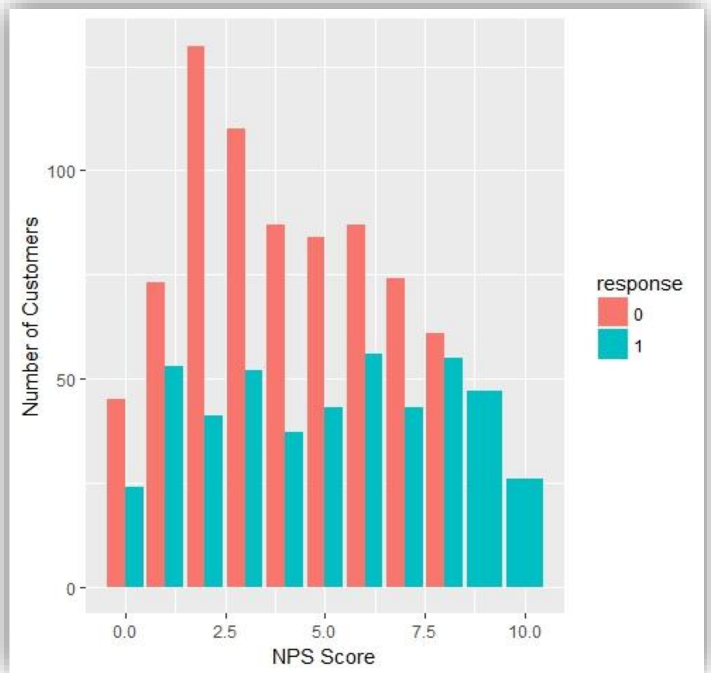
Visualizing Distribution Graphically For Significant Predictors

NPS Score by Response to the Campaign

Boxplot for NPS By Response to the Campaign



Stacked Bar Chart(Customers By Response to the Campaign and NPS Score)



- The Customers who have mentioned NPS score as 9/10 ,have all responded to the campaign which indicates higher satisfaction level has a positive relationship with Response to the Campaign
- The Customers whose response is 1 seem to have increasing trend till 6 years post which we can see a slight drop

Logistic Regression : R code for Odds Ratio

```
fmcg_glmmodel <- glm(response ~
Annualsales+sales_Q4+Sales_B1+Pcontri_B1+n_yrs+buyingfreq+buyingfreq_B1+Region+n
ps+loyalty+portal+n_comp+email+sms+call+rewards+brandengagement+growth,
data=Master_FMCG, family=binomial)

coef(fmcg_glmmodel)

exp(coef(fmcg_glmmodel))

cbind(odds_ratio = exp(coef(fmcg_glmmodel)),exp(confint(fmcg_glmmodel)))
```

Odds Ratio :Interpretation

	Estimate	Odds_ratio	Interpretation
(Intercept)	-2.80301	0.06062754	
n_yrs	0.09106	1.09533604	For one unit change in n_yrs, the odds of response to the campaign will increase by 1.096 years
Nps	0.15765	1.17075926	For one unit change in nps ,the odds of the response to the campagin will increase by 1.18 units
Email	0.63788	1.8924667	For one unit change in email ,the odds of the response to the campaign will increase by 1.89 units
sms	0.41289	1.51117639	For one unit change in sms ,the odds of the response to the campaign will increase by 1.51 units
call	0.62035	1.85957625	For one unit change in call ,the odds of the response to the campaign will increase by 1.86 units

Logistic Regression :Classification Table

Cut off value	**Sensitivity**	**Specificity**
0.5	42%	84%
0.4	60%	67%
0.39	62%	65%
0.37	66%	61%

Sensitivity : % of occurrences correctly predicted

Sensitivity : % of non occurrences correctly predicted

* 0.39 can be considered as the optimum cut off value

Logistic Regression: Hold – Out Cross Validation

- In Hold out Cross-validation method ,the data was split into 2 non overlapping parts: ’Training Data’ and ‘Testing Data’
- The model was developed using training data by taking 80% of the total sample and evaluated using testing data using remaining 20% of the sample
- Cross validation results were evaluated using Confusion Matrix
- ROC curve was generated first for training data and then for Testing data. Area under the curve measured using auc value for both training and testing data sets.

Logistic Regression: Hold – Out Cross Validation

```
library(caret)
index <- createDataPartition(Master_FMCG$response,p=0.8,list=F)

traindata <- Master_FMCG[index,]
testdata <- Master_FMCG[-index,]

traindata$predprob <- predict(fmcg_glmmodel,traindata,type='response')
traindata$predY <- ifelse(traindata$predprob>0.39,1,0)

confusionMatrix(traindata$predY,traindata$response,positive = "1")

traindata$predprob <- predict(fmcg_glmmodel,traindata,type='response')
pred <- prediction(traindata$predprob,traindata$response)

perf <- performance(pred,"tpr","fpr")

plot(perf)
abline(0,1)

auc <- performance(pred,"auc")
```

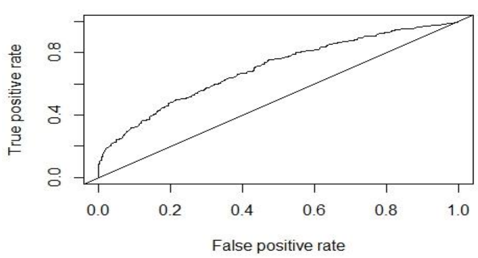
Hold – Out Cross Validation

Confusion Matrix Statistics

	Reference	
Prediction	0	1
0	398	139
1	206	240

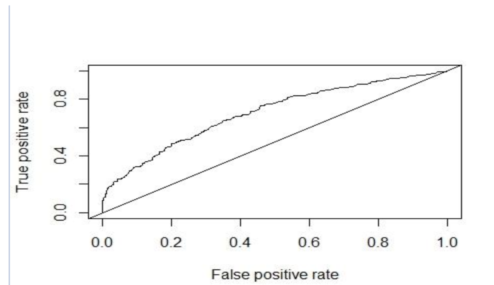
Accuracy : 0.649
95% CI : (0.6183, 0.6789)
No Information Rate : 0.6144
P-Value [Acc > NIR] : 0.0137029
Kappa : 0.2829
Mcnemar's Test P-Value : 0.0003804
Sensitivity : 0.6332
Specificity : 0.6589
Pos Pred Value : 0.5381
Neg Pred Value : 0.7412
Prevalence : 0.3856
Detection Rate : 0.2442
Detection Prevalence : 0.4537
Balanced Accuracy : 0.6461
'Positive' Class : 1

Logistic Regression :ROC Curve in R



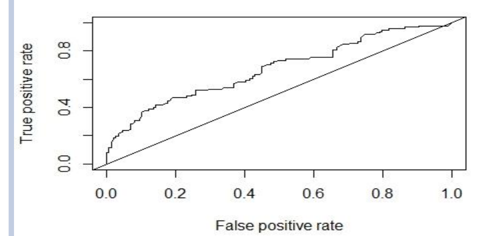
Model(N=1228)

Area under the ROC Curve=69%



Training Data(N=983)

Area under the ROC Curve=70%



Test Data(N=245)

Area under the ROC Curve=67%

K-Fold Cross Validation

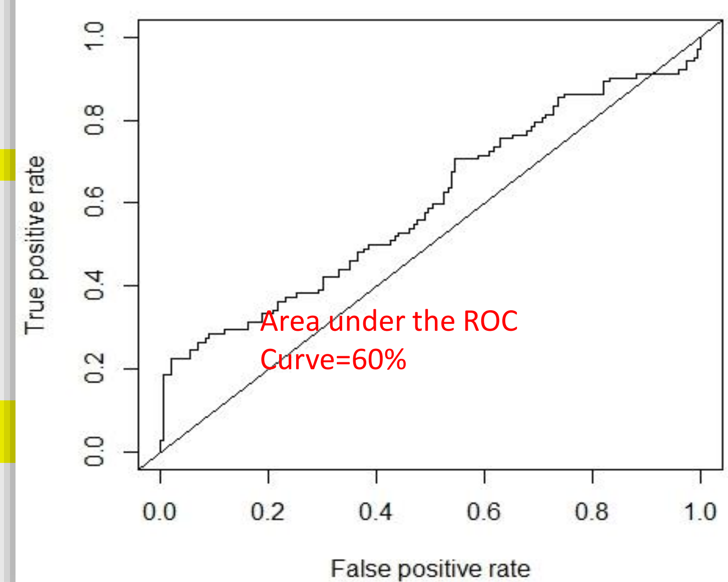
When evaluating models, we often want to assess how well it performs in predicting the target variable on different subsets of the data. One such technique for doing this is k-fold cross-validation, which partitions the data into k equally sized segments (called ‘folds’).

One fold is held out for validation while the other k-1 folds are used to train the model and then used to predict the target variable in our testing data.

This process is repeated k times, with the performance of each model in predicting the hold-out set being tracked using a performance metric such as accuracy. We have used the most common variation of cross validation that is 10-fold cross-validation.

Logistic Regression :K-Fold Cross Validation

```
> ctrl <- trainControl(method = "repeatedcv", number = 10, savePredictions = TRUE)
> glmmod_fit <- train(response ~ Annualsales+sales_Q4+Sales_B1+Pcontri_B1+n_yrs+
+                     buyingfreq+buyingfreq_B1+Region+nps+loyalty+portal+n_comp+email+
+                     sms+call+rewards+brandengagement+growth,data=traindata, method="glm", family="binomial",
+                  trControl = ctrl, tuneLength = 5)
> predg <- predict(glmmod_fit, newdata=testdata)
> confusionMatrix(predg, testdata$response)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 117  68
         1  26  34

               Accuracy : 0.6163
                 95% CI : (0.5523, 0.6775)
    No Information Rate : 0.5837
    P-Value [Acc > NIR] : 0.1656

                  Kappa : 0.161
 Mcnemar's Test P-Value : 2.349e-05

            Sensitivity : 0.8182
            Specificity : 0.3333
         Pos Pred Value : 0.6324
         Neg Pred Value : 0.5667
             Prevalence : 0.5837
         Detection Rate : 0.4776
   Detection Prevalence : 0.7551
      Balanced Accuracy : 0.5758
```



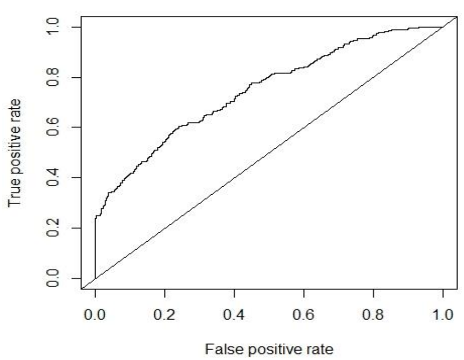
NAÏVE BAYES

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.6115635 0.3884365

Conditional probabilities:
   Annualsales
Y       [,1]     [,2]
  0 54515.15 42844.32
  1 53516.95 42997.44

   sales_Q4
Y       [,1]     [,2]
  0 13926.08 20516.42
  1 13141.11 19004.57

   Sales_B1
Y       [,1]     [,2]
  0 11080.03 18140.52
  1 10813.49 18439.05

   Pcontri_B1
```

Output shows a list of tables ,one for each predictor variables
In this data, all predictor variables are numeric except Region, for numeric predictors, the output is shown for each target class with mean and standard deviation.
For example :For 'Annual sales', mean of churn status =0 is 54515 and SD is 42844
Mean of churn status =1 is 53517 and SD is 42997

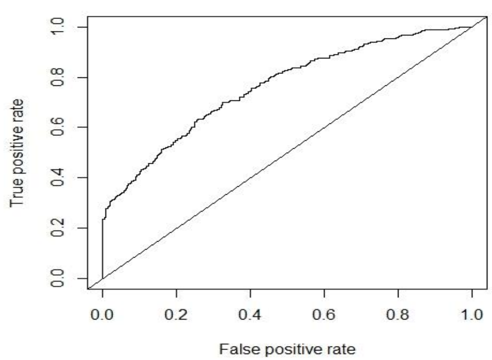
NAÏVE BAYES

Hold Out Cross Validation: Area Under the ROC Curve



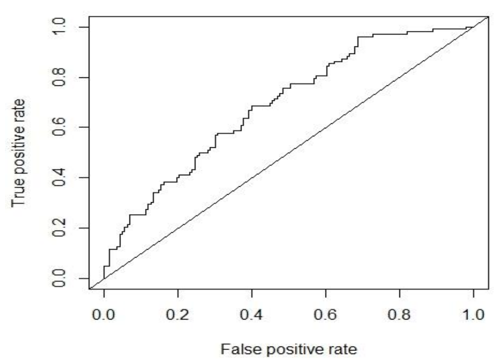
Model(N=1228)

Area under the ROC Curve=71%



Training Data(N=983)

Area under the ROC Curve=72%



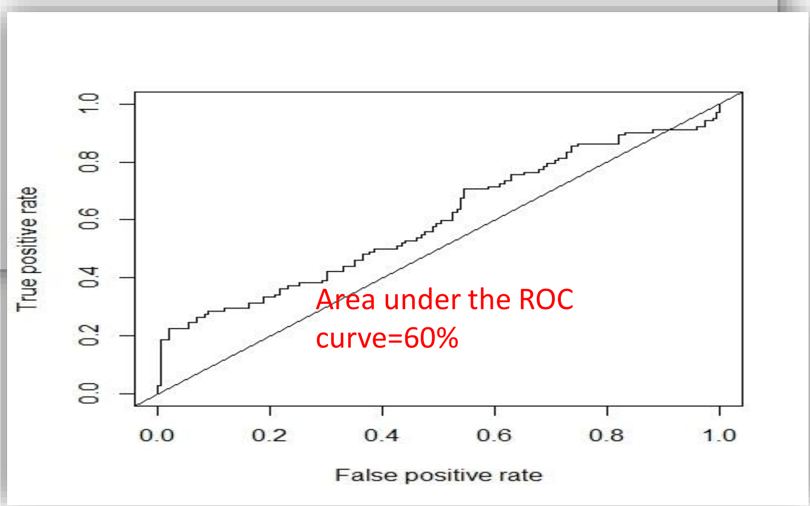
Test Data(N=245)

Area under the ROC Curve=69%

NAÏVE BAYES

K fold Cross Validation: Area Under the ROC Curve and Confusion Matrix

```
> kfolds<-trainControl(method="cv",number=10)
> naivemodel <- train(response ~ Annualsales+sales_Q4+Sales_B1+Pcontri_B1+n_yrs+
+
+                  buyingfreq+buyingfreq_B1+nps+loyalty+portal+n_comp+email+
+
+                  sms+call+rewards+brandengagement+growth,data=traindata,
+
+                method="nb",trControl=kfolds)
There were 50 or more warnings (use warnings() to see the first 50)
> predk_n<-as.data.frame(predict(naivemodel$finalModel,testdata))
There were 50 or more warnings (use warnings() to see the first 50)
> head(predk)
           0          1
4   0.6933333 0.30666667
14  0.9033333 0.09666667
15  0.1900000 0.81000000
16  0.7333333 0.26666667
19  0.7166667 0.28333333
23  0.1833333 0.81666667
```

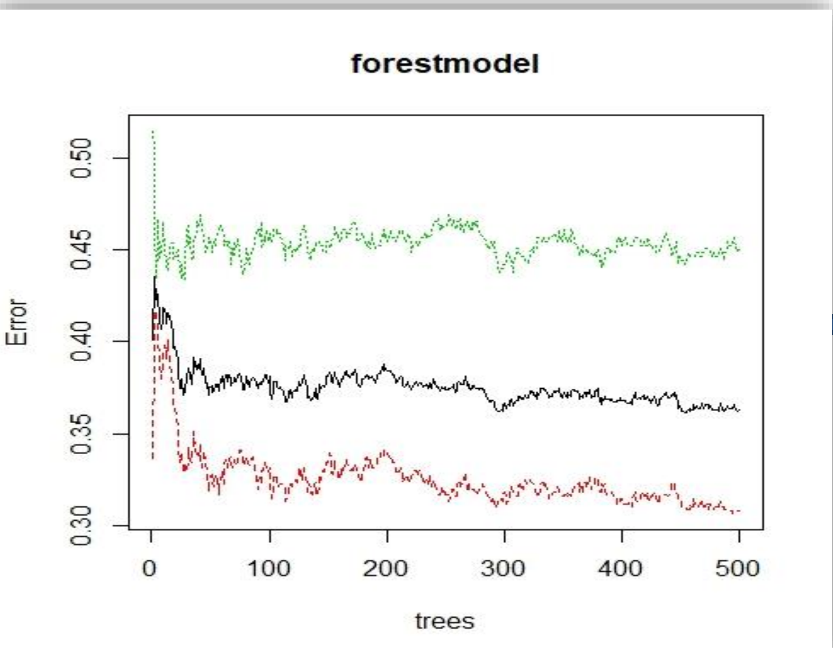



```
> confusionMatrix(testdata$predY_n,testdata$response)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 125  72
         1  18  30

               Accuracy : 0.6327
                 95% CI : (0.5689, 0.6931)
    No Information Rate : 0.5837
    P-Value [Acc > NIR] : 0.06738

                  Kappa : 0.1821
 Mcnemar's Test P-Value : 2.314e-08

            Sensitivity : 0.8741
            Specificity : 0.2941
         Pos Pred Value : 0.6345
         Neg Pred Value : 0.6250
             Prevalence : 0.5837
         Detection Rate : 0.5102
   Detection Prevalence : 0.8041
      Balanced Accuracy : 0.5841

       'Positive' Class : 0
```

Random Forest

Model Output

```
Call:
 randomForest(formula = response ~ Annualsales + sales_Q4 + Sales_B1 +Pcontri_B1 + n_yrs + buyingfreq + buyingfreq_B1 + Region +nps + loyalty + portal + n_comp + email +
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 36.32%
Confusion matrix:
    0   1 class.error
0 520 231   0.3075899
1 215 262   0.4507338
```



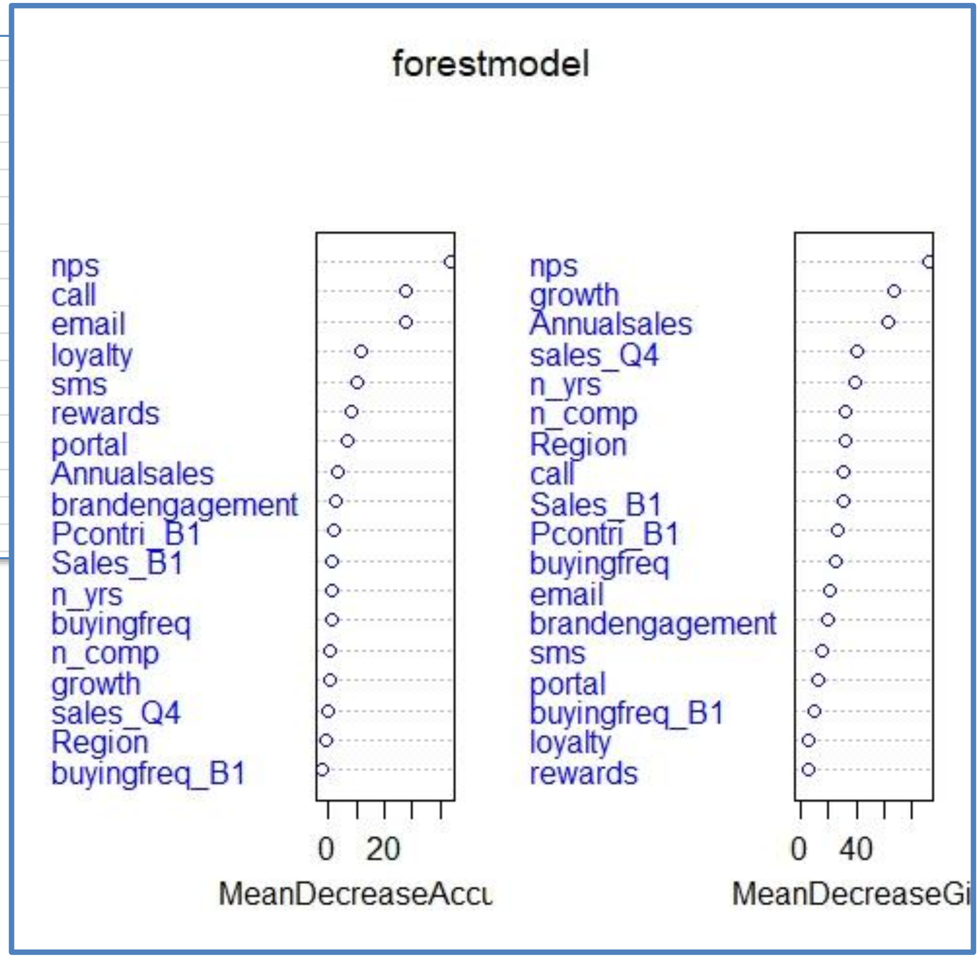
Decision Trees Error Rate

- 500 decision trees /forest has been built using the Random Forest algorithm. Plot shows error rate for all 500 decision trees. Black line shows the overall OOB error rate. Coloured lines shows error rate for each class.

Random Forest

Variable Importance

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Annualsales	0.0118164048	-1.104492e-02	0.0028743321	62.946836
sales_Q4	0.0018505063	-3.752332e-03	-0.0002893309	40.187492
Sales_B1	0.0046209969	-4.821447e-03	0.0009132824	30.238649
Pcontri_B1	0.0043783107	-4.227900e-03	0.0010446639	26.864873
n_yrs	0.0014734869	-6.810409e-06	0.0008942593	39.522184
buyingfreq	0.0032484857	-3.264446e-03	0.0007223739	25.492899
buyingfreq_B1	0.0002071338	-2.596178e-03	-0.0008676920	9.800664
Region	-0.0009003029	-3.173748e-04	-0.0006376872	31.888576
nps	0.0359492228	4.521121e-02	0.0395055588	91.626364
loyalty	0.0210673062	-1.292600e-02	0.0078828286	5.987817
portal	0.0062008046	-2.219143e-03	0.0029368909	12.333804
n_comp	0.0011522252	-9.142321e-04	0.0003672096	32.661797
email	0.0622853859	-1.383736e-02	0.0327259020	21.535539
sms	0.0102660751	-3.017321e-03	0.0051139929	15.744947
call	0.0360831326	3.872419e-03	0.0236200308	30.717885
rewards	0.0144178622	-8.657917e-03	0.0054797937	5.019697
brandengagement	0.0059373931	-5.778997e-03	0.0013901620	19.187611
growth	0.0022169254	-3.314638e-03	0.0001090362	67.420708

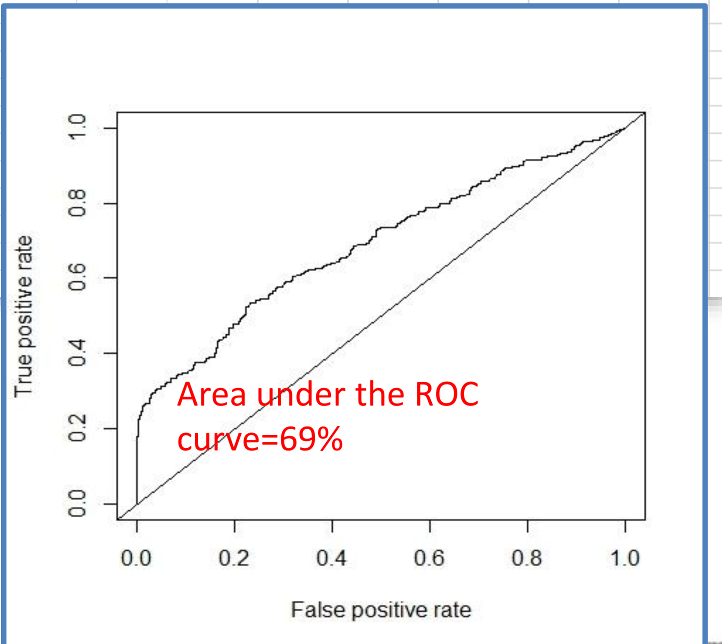


- Based on Random Forest Variable Importance plot, nps score is the most important variable.
- Email and Call seem to be the second and third most important variables with a slight difference and hence can conclude that both of them are the most effective communication channels.

Random Forest

Training Data :Model Output and Area under ROC Curve

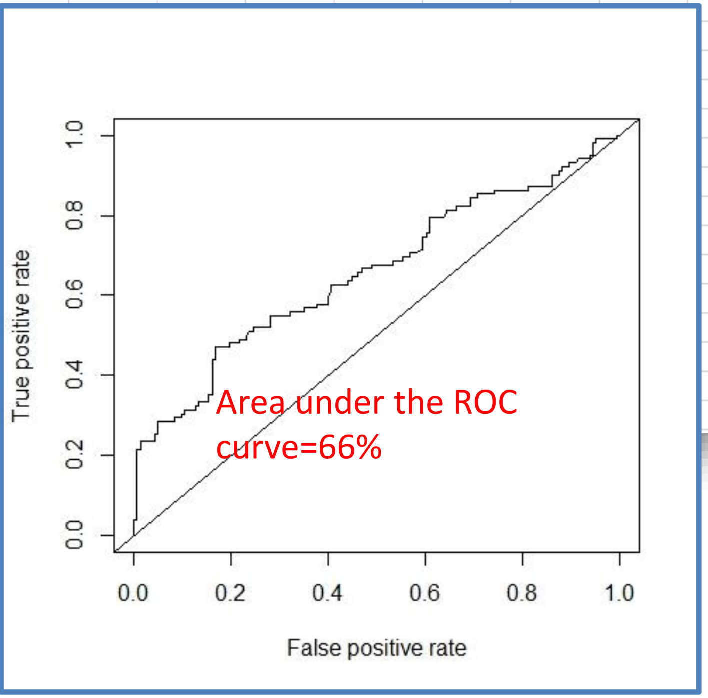
```
Call:
 randomForest(formula = response ~ Annualsales + sales_Q4 + Sales_B1 +Pcontri_B1 + n_yrs + buyingfreq + buyingfreq_B1 + Region +nps + loyalty + portal + n_comp + email +
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 34.18%
Confusion matrix:
    0   1 class.error
0 431 177   0.2911184
1 159 216   0.4240000
> predtree <- predict(forestmodel_train,traindata,type="prob")
> pred <- prediction(forestmodel_train$votes[,2],traindata$response)
> perf <-performance(pred,"tpr","fpr")
> plot(perf)
> abline(0,1)
> auc <- performance(pred,"auc")
> auc@y.values
[[1]]
[1] 0.6891601
```



Random Forest

Test Data :Model Output and Area under ROC Curve

```
Call:
 randomForest(formula = response ~ Annualsales + sales_Q4 + Sales_B1 +Pcontri_B1 + n_yrs + buyingfreq + buyingfreq_B1 + Region +nps + loyalty + portal + n_comp + email +
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 39.59%
Confusion matrix:
   0  1 class.error
0 84 59   0.4125874
1 38 64   0.3725490
> predtree <- predict(forestmodel_test,testdata,type="prob")
>
> pred <- prediction(forestmodel_test$votes[,2],testdata$response)
> perf <-performance(pred,"tpr","fpr")
> plot(perf)
> abline(0,1)
> auc <- performance(pred,"auc")
> auc@y.values
[[1]]
[1] 0.6566571
```

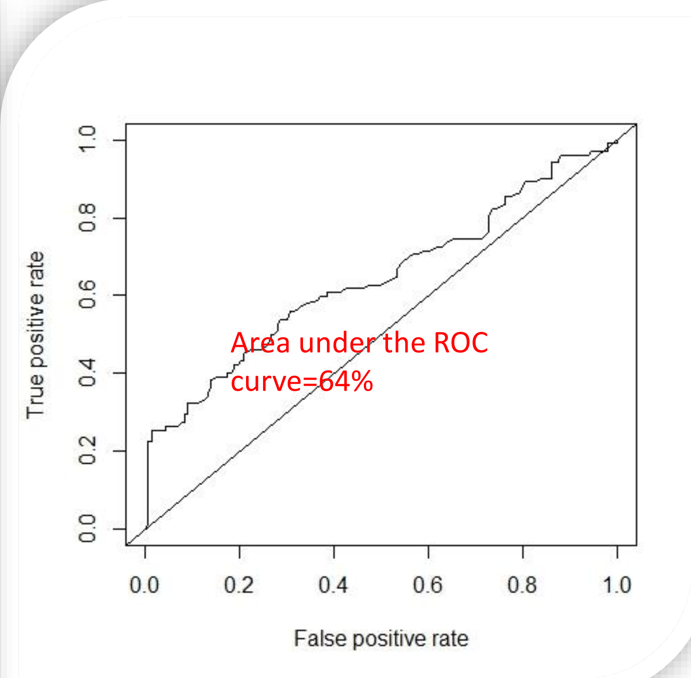


Random Forest

K-fold Cross Validation :Predicted Probabilites, Confusion Matrix and Area Under ROC Curve

```
> predk <- predict(fit_rf_train,testdata,type="prob")
> predk
           0          1
4   0.6933333 0.30666667
14  0.9033333 0.09666667
15  0.1900000 0.81000000
16  0.7333333 0.26666667
19  0.7166667 0.28333333
23  0.1833333 0.81666667
25  0.3266667 0.67333333
28  0.2033333 0.79666667
29  0.7300000 0.27000000
34  0.5833333 0.41666667
35  0.6733333 0.32666667
43  0.7400000 0.26000000
44  0.4633333 0.53666667
```

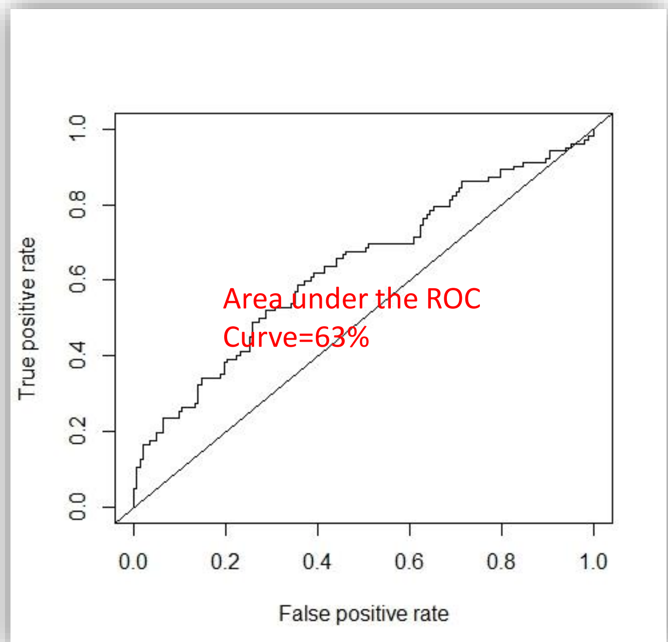
```
> confusionMatrix(testdata$predYr,testdata$response)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 111  55
         1  32  47

               Accuracy : 0.6449
                 95% CI : (0.5815, 0.7048)
    No Information Rate : 0.5837
    P-Value [Acc > NIR] : 0.02936

                  Kappa : 0.2449
 Mcnemar's Test P-Value : 0.01834

            Sensitivity : 0.7762
            Specificity : 0.4608
         Pos Pred Value : 0.6687
         Neg Pred Value : 0.5949
             Prevalence : 0.5837
         Detection Rate : 0.4531
   Detection Prevalence : 0.6776
      Balanced Accuracy : 0.6185

       'Positive' Class : 0
```



Support Vector Machines

Hold Out Cross Validation: Confusion Matrix and Area under the ROC Curve

```
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.04761905

Number of Support Vectors:  656

> predsvm <- predict(model_svm,testdata)
> confusionMatrix(predsvm,testdata$response)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 128  76
         1  15  26

               Accuracy : 0.6286
                 95% CI : (0.5648, 0.6892)
    No Information Rate : 0.5837
    P-Value [Acc > NIR] : 0.08622

                  Kappa : 0.1641
 Mcnemar's Test P-Value : 3.181e-10

            Sensitivity : 0.8951
            Specificity : 0.2549
         Pos Pred Value : 0.6275
         Neg Pred Value : 0.6341
             Prevalence : 0.5837
         Detection Rate : 0.5224
   Detection Prevalence : 0.8327
      Balanced Accuracy : 0.5750

       'Positive' Class : 0
```



Best Model?

Best Model Selection

Method	Cross-Validation Method	Accuracy	Sensitivity	Specificity	AUC
Binary Logistic Regression	K-Fold	61%	81%	33%	60%
Naïve Bayes Classifier	K-Fold	63%	88%	29%	60%
Random Forest	K-Fold	64%	78%	46%	64%
Support Vector Machines	Hold Out	62%	90%	25%	63%

Since Random Forests gives Highest AUC compared to other models, we will go with Random Forests Model as the Final one and Calculate predicted probabilities based on this model.

Random Forest model can be used to select a targeted base for the next campaign

Thank You!